SYSTEM AND METHOD FOR DISTRIBUTED WEB CRAWLING

## ABSTRACT OF THE DISCLOSURE

The present invention provides for the efficient downloading of data set addresses from among a plurality of host computers, using a plurality of web crawlers. Each web crawler identifies URL's in data sets downloaded by that web crawler, and identifies the host computer identifier within each such URL. The host computer identifier for each URL is mapped to the web crawler identifier of one of the web crawlers. If the URL is mapped to the web crawler identifier of a different web crawler, the URL is sent to that web crawler for processing, and otherwise the URL is processed by the web crawler that identified the URL. Each web crawler sends URL's to the other web crawlers for processing, and each web crawler receives URL's from the other web crawlers for processing. In a preferred embodiment, each web crawler processes only the URL's assigned to it, which are the URL's whose host identifier is mapped to the web crawler identifier for that web crawler. Each web crawler filters the URL's assigned to it by comparing them against a database of URL's already known by the web crawler and removing the already known URL's. If a URL is not already known to the web crawler, the data set corresponding to the URL is scheduled for downloading.